

Parcial 1 TAM

1. Sea el modelo de regresión $t_n = \phi(x_n)W^T + \eta_n$, con $t_n \in \mathbb{R}$, $x_n \in \mathbb{R}^p$, $W \in \mathbb{R}^{q \times p}$, $\phi: \mathbb{R}^p \rightarrow \mathbb{R}^q$, $q > p$ y $\eta_n \sim N(\eta_n | 0, \sigma^2)$. Presente el problema de optimización (inferencial) la solución del mismo para los modelos mínimos cuadrados, mínimos cuadrados regularizados, máxima verosimilitud, máximo a posteriori y bayesiano con modelo lineal Gaussiano. Aunque datos iid. Discuta las diferencias y similitudes entre los modelos estudiado.

Tenemos el modelo de regresión

$$t_n = \phi(x_n)W^T + \eta_n \text{ donde}$$

$$\begin{aligned} t_n &\in \mathbb{R} \\ x_n &\in \mathbb{R}^p \\ W &\in \mathbb{R}^{q \times p} \\ \phi: \mathbb{R}^p &\rightarrow \mathbb{R}^q \\ \eta_n &\sim N(\eta_n | 0, \sigma^2) \end{aligned}$$

1.1 Mínimos cuadrados ordinarios

Es una técnica usada para encontrar el mejor ajuste de una función a un conjunto de datos encontrando los parámetros de una función lineal que minimicen la suma de los cuadrados de las diferencias entre los valores observados y los valores predichos por la función.

$$\hat{W} = \frac{1}{n} \sum_{i=1}^n (t_i - \phi(x_i)W^T)^2, \text{ ahora de forma matricial:}$$

$$\hat{t} = f(\mathbf{x} | W) = \begin{bmatrix} \hat{t}_1 \\ \hat{t}_2 \\ \vdots \\ \hat{t}_n \end{bmatrix} = \phi W^T ; \quad t = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_n \end{bmatrix}$$

Queremos encontrar el arg que minimice la diferencia.

$$\boxed{\hat{W} = \arg \min_W \| t - \phi W^T \|_2^2} \rightarrow \text{función costo con min cuadrados para regresión lineal.}$$

ahora derivamos, e igualamos a 0:

$$\frac{d \|t - \phi w^T\|_2^2}{dw} = 0 = \frac{d \|e\|_2^2}{dw} = 0$$

$$\frac{d \langle t - \phi w^T, t - \phi w^T \rangle}{dw} = 0$$

$$\frac{d (t - \phi w^T)^T (t - \phi w^T)}{dw} = 0$$

$$\frac{d (t^T - w^T \phi^T) (t - \phi w^T)}{dw} = 0$$

$$\frac{d}{dw} t^T t - t^T \phi w^T - w^T \phi^T t + w^T \phi \phi w^T = 0$$

$$\frac{d}{dw} t^T t - 2 t^T \phi w^T + w^T \phi^T \phi w^T = 0$$

$$\frac{d \{ \dots \}}{dw} = 0 - 2 t^T \phi + 2 w^T \phi^T \phi = 0$$

$$2 w^T \phi^T \phi = 2 t^T \phi$$

$$\hat{w} = t^T \phi (\phi^T \phi)^{-1}$$

Optimización de los w

1.2. Mínimos Cuadrados Regularizados.

la idea de mínimos cuadrados regularizados es agregar un término de penalización a la función de costo original, con el objetivo de controlar el tamaño de los coeficientes de regresión. Así minimizando la función de costo ajustando los pesos w de manera que tanto el error de predicción como la magnitud de los coeficientes sean pequeños.

El término $\lambda \|w\|_2^2$ penaliza los valores grandes de los coeficientes, lo que ayuda a prevenir el sobreajuste al restringir la complejidad del modelo.

Además el término $\lambda \|w\|_2^2$ al ser agregado al modelo de mínimos cuadrados nos ayuda a garantizar que la matriz que se debe invertir sea invertible.

Cuando la matriz $(X^T X)^{-1}$ no es de rango completo o está cerca de ser singular, agregar el término $\lambda \|w\|_2^2$ ayuda a estabilizar la inversión de la matriz, lo que facilita la solución del problema de optimización de flujos y mejora la robustez del modelo.

Tenemos la tarea de regresión:

$$t_n = \phi w^T + \eta_n.$$

Ahora tenemos el modelo para la función de costo con MCR.

$$w = \arg \min_w (||t - \phi w^T||_2^2 + \lambda \|w\|_2^2).$$

buscamos los valores mínimos derivando e igualando a 0.

$$\frac{d}{dW} \|t - \Phi W^T\|^2 + \lambda \|W\|^2 = 0$$

$$dW$$

$$\frac{d}{dW} \langle t - \Phi W^T, t - \Phi W^T \rangle + \lambda \langle W, W \rangle = 0$$

$$dW$$

$$\frac{d}{dW} (t^T - \Phi W^T)^T (t^T - \Phi W^T) + \lambda ((W^T)^T (W)) = 0$$

$$dW$$

$$\frac{d}{dW} (t^T t - t^T \Phi W^T - W^T \Phi^T t + W^T \Phi^T \Phi W^T + \lambda W^T W) = 0$$

$$dW$$

$$\frac{d}{dW} (t^T t - 2t^T \Phi W^T + W^T \Phi^T \Phi W^T + \lambda W^T W) = 0$$

$$dW$$

$$\frac{d}{dW} [\dots] = 0 - 2t^T \Phi + 2W \Phi^T \Phi + 2\lambda W = 0$$

$$2W \Phi^T \Phi + 2\lambda W = 2t^T \Phi$$

$$W (\Phi^T \Phi + \lambda I) = t^T \Phi$$

$$W = t^T \Phi (\Phi^T \Phi + \lambda I)^{-1}$$

Solución para minimos cuadrados regulados

1.3. Máxima verosimilitud.

La función máxima verosimilitud se define como la probabilidad de observar los datos de entrada x_n , asumiendo datos iid y una distribución gaussiana para el ruido.

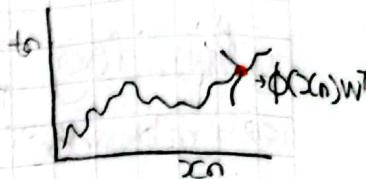
• Asumimos datos iid $\rightarrow x_n \sim N(x_n | \mu, \sigma^2)$.

Tenemos la tarea de regresión

$$t_n = \phi(x_n)W^T + \eta_n;$$

$$\eta_n \sim N(\eta | 0, \sigma^2).$$

$$\eta_n = t_n - \phi(x_n)W^T \sim N(t_n | \phi(x_n)W^T, \sigma^2).$$



Lo que buscamos es maximizar la probabilidad de la normal $N(t_n | \phi(x_n)W^T, \sigma^2)$.

$$\text{maximizar} \rightarrow P(t_n | \phi(x_n)W^T, \sigma^2)$$

Lo que es igual a maximizar la función normal $N(t_n | \phi(x_n)W^T, \sigma^2)$.

$$P(t_n | \phi(x_n)W^T, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_n - \phi(x_n)W^T)^2}{2\sigma^2}\right) = N(t_n | \phi(x_n)W^T, \sigma^2)$$

Ahora sabiendo cuál es nuestra función objetivo para estimar los pesos W , definir nuestra función de verosimilitud $L(W)$, la cual representa la probabilidad conjunta de observar todos los datos t_n dados los x_n y el peso W .

Recordemos que: $P(X) = p(x_1, x_2, \dots, x_n) = P(x_1) \cdot P(x_2) \cdots P(x_n)$

Lo que es igual en términos de normales:

$$= N(x_1 | \mu, \sigma^2) \cdot N(x_2 | \mu, \sigma^2) \cdots N(x_n | \mu, \sigma^2)$$

Por tanto esto lo podemos expresar como una multiplicación.

$$P(X) = \prod_{n=1}^N N(x_n | \mu, \sigma^2) \rightarrow \text{Para nuestro caso, asumimos una normal.}$$

Ahora para simplificar los cálculos en busca del argumento max de w , vamos a implementar la función logaritmo ya que esta nos ofrece ventajas como transformar productos en sumas, además al trabajar con probabilidades pequeñas que tienden a cero puede causar problemas esto también se minimiza al cambiar productos por sumas de logaritmos.

Entonces aplicando log a la multiplicatividad tenemos:

$$\log f(w) = \log \prod_{n=1}^N N(t_n | \phi(x_n)w^T, \sigma^2)$$

reescibiendo

$$\log \prod_{n=1}^N \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(t_n - \phi(x_n)w^T)^2}{2\sigma^2}\right) = \log L(w).$$

por propiedades de los logaritmos podemos reescribir la multiplicatividad como una suma

$$\log\left(\prod_{i=1}^N a_i\right) = \sum_{i=1}^N \log(a_i)$$

$$\sum_{n=1}^N \log\left(\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(t_n - \phi(x_n)w^T)^2\right)\right)$$

ahora aplicamos $\log(ab) = \log(a) + \log(b)$

$$\sum_{n=1}^N \left[\log\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right) + \log\left(\exp\left(-\frac{1}{2\sigma^2}(t_n - \phi(x_n)w^T)^2\right)\right) \right]$$

$$\text{ahora } \log(\exp(c)) = C \quad y \quad \log\left(\frac{1}{x}\right) = -\log(x)$$

$$\sum_{n=1}^N \left[-\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} (t_n - \phi(x_n)w^T)^2 \right] = \log L(w)$$

$$-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{n=1}^N (t_n - \phi(x_n)w^T)^2 = \log L(w)$$

ahora de forma matricial tenemos

$$-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t - \Phi W^T\|_2^2$$

queremos encontrar el argumento que maximice
nuestra probabilidad.

$$W = \arg \max_W -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t - \Phi W^T\|_2^2$$

derivamos e igualamos a cero

$$\frac{d}{dW} \left[-\frac{N}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \|t - \Phi W^T\|_2^2 \right] = 0$$

sabemos que la derivada de $-\frac{N}{2} \log(2\pi\sigma^2)$ va ser
igual a 0 ya que es constante. quedando:

$$\frac{d}{dW} \left[-\frac{1}{2\sigma^2} \langle t - \Phi W^T, t - \Phi W^T \rangle \right] = 0$$

$$\frac{d}{dW} \left[\frac{1}{2\sigma^2} (t - \Phi W^T)^T (t - \Phi W^T) \right] = 0$$

$$\frac{d}{dW} \left[-\frac{t^T t}{2\sigma^2} + \frac{2t^T \Phi W^T}{2\sigma^2} - \frac{W \Phi^T \Phi W^T}{2\sigma^2} \right] = 0$$

$$\frac{d}{dW} \left\{ \dots \right\} = 0 + \frac{t^T \Phi}{\sigma^2} - \frac{2W \Phi^T \Phi}{2\sigma^2} = 0$$

$$\frac{d}{dW} \left\{ \dots \right\} = \frac{t^T \Phi}{\sigma^2} - \frac{W \Phi^T \Phi}{\sigma^2} = 0, \quad \frac{t^T \Phi}{\sigma^2} = \frac{W \Phi^T \Phi}{\sigma^2}$$

$$\frac{d}{dW} \left\{ \dots \right\} = \boxed{\frac{t^T \Phi}{\sigma^2} (\Phi^T \Phi)^{-1} = W}$$

2.4. Máximo a posteriori (MAP).

Partimos de que tenemos información previa o conocimiento previo sobre los parámetros del modelo, representado por una distribución de probabilidad a priori. Este enfoque combina la información de los datos observados (likelihood) con la información a priori para obtener una estimación más precisa de los parámetros del modelo.

En máx a posteriori partimos de un conjunto de datos observados y una distribución de probabilidad previa sobre los parámetros.

Iniciamos con nuestro modelo de regresión

$$t_n = \phi(x_n)w^T + \eta_n; \quad \eta_n \sim N(\eta | 0, \sigma_n^2)$$

$$\eta_n = t_n - \phi(x_n)w^T \sim N(t_n | \phi(x_n)w^T, \sigma^2) \rightarrow \text{likelihood}$$

Además, asumimos una distribución sobre los parámetros:

$$P(w) \sim N(w | 0, \sigma_w^2) \rightarrow \text{Prior}$$

Por bayes tenemos las siguientes relaciones:

entre la distribución a priori $P(w)$, la verosimilitud $P(t_n | \phi(x_n)w, \sigma^2)$ y la distribución a posteriori $P(w | t_n, \phi(x_n)w^T, \sigma^2)$.

esta relación por bayes se expresa como:

$$P(w | t_n, \phi(x_n), \sigma^2) = \frac{P(t_n | \phi(x_n)w^T, \sigma^2) \cdot P(w)}{P(t_n | \phi(x_n))}$$

prior
likelihood
evidencia

Entonces la idea es maximizar la distribución a posteriori $P(w | t_n, \phi(x_n))$, que representa la probabilidad de los parámetros w dados los datos t_n y las características $\phi(x_n)$. Esto es equivalente a maximizar

el producto de la verosimilitud $P(t_n | \phi(x_n), W)$ y la distribución a priori $P(W)$, ya que la evidencia $P(t_n | \phi(x_n))$ es constante respecto a W dando como resultado:

$$P(W | t_n, \phi(x_n)) \propto P(t_n | \phi(x_n), W) \cdot P(W).$$

Ahora vamos a maximizar la log-distribución a posteriori, esto es análogo a minimizar la función de costo en el enfoque de regresión regularizada, donde la log-verosimilitud representa el ajuste a los datos y la log-prior actúa como una penalización que incorpora conocimiento previo o restricciones sobre los parámetros W .

$$W_{MAP} = \arg \max_W \log \left(\prod_{n=1}^N N(t_n | \phi(x_n) W^T, \sigma_n^2) \cdot \prod_{q=1}^Q N(W_q | 0, \sigma_w^2) \right)$$

$$W_{MAP} = \arg \max_W -\frac{1}{2\sigma_n^2} \|t - \phi W^T\|_2^2 - \frac{1}{2\sigma_w^2} \|W\|_2^2$$

Podemos factorizar el problema como:

$$W_{MAP} = \arg \min_W \|t - \phi W^T\|_2^2 + \frac{\sigma_n^2}{\sigma_w^2} \|W\|_2^2; \quad \lambda = \frac{\sigma_n^2}{\sigma_w^2}$$

Siendo equivalente con mínimos cuadrados regularizado y teniendo en cuenta la demostración de Max likelihood llegamos a:

$$W = t^T \phi (\phi^T \phi + \lambda \mathbb{I})^{-1}$$

1.5. Bayesiano con modelo lineal Gaussiano.

Sea el prior $P(x) = N(x | \mu, \Lambda^{-1})$

ademas, sea la verosimilitud. desde un modelo lineal $y = Ax + b$

$$P(y|x) = N(y | Ax + b, L^{-1})$$

completando cuadradecos sobre la gaussiana conjunta, tenemos que:

$$P(y) = N(y | A\mu + b, L^{-1} + A\Lambda^{-1}A^T)$$

$$P(x|y) = N(y | \mu_{x|y}, \Sigma_{x|y})$$

con:

$$\mu_{x|y} = \Sigma_{x|y} (A^T L (y - b) + \Lambda \mu)$$

$$\Sigma_{x|y} = (\Lambda + A^T L A)^{-1}$$

Entonces para el modelo de regresión

$$t_n = \phi(x_n) W^T + \eta_n$$

$$\eta_n \sim P(\eta_n) = N(\eta_n | 0, G_n^2)$$

$$\eta_n = t_n - \phi(x_n) W^T$$

Por tanto la verosimilitud se puede escribir como:

$$p(t_n | \phi(x_n) W^T, \sigma_n^2) = N(t_n | \phi(x_n) W^T, \sigma_n^2)$$

En forma vectorial:

$$p(t | \phi W^T, \sigma_n^2) = N(t | \phi W^T, \sigma_n^2)$$

Asumiendo el prior.

$$p(W) = N(W | M_0, S_0)$$

El posterior se puede estimar como:

$$p(W | t) = N(W | M_N, S_N)$$

Donde:

$$M_N = S_N^{-1} M_0 + \frac{1}{\sigma_n^2} \phi^T t$$

$$S_N = (S_0^{-1} + \frac{1}{\sigma_n^2} \phi \phi^T)^{-1}$$

Si se impone un prior de la forma:

$$p(W) = N(W | \emptyset, \sigma_W^2)$$

Entonces:

$$p(W | t) = N(W | \tilde{M}_N, \tilde{S}_N)$$

$$\tilde{M}_N = \frac{1}{\sigma_n^2} \tilde{S}_N \phi^T t$$

$$\tilde{S}_N = \left(\frac{1}{\sigma_n^2} I_Q + \frac{1}{\sigma_n^2} \phi^T \phi \right)^{-1} = \left(\frac{1}{\sigma_n^2} \right) \left(\frac{\sigma_n^2}{\sigma_W^2} I_Q + \phi \phi^T \right)^{-1}$$

Reemplazando en la media condicional

$$\tilde{m}_N = \frac{1}{\sigma_n^2} \left(\frac{1}{\sigma_n^2} \right)^{-1} \left(\frac{\sigma_n^2}{\sigma_w^2} I_Q + \phi^\top \phi \right)^{-1} \phi^\top t$$

$$\tilde{m}_N = \left(\frac{\sigma_n^2}{\sigma_w^2} I_Q + \phi^\top \phi \right)^{-1} \phi^\top t$$

Nota: la solución del modelo lineal Gaussiano para el prior $p(w) = N(w | 0, G_w)$ y ante ruido blanco Gaussiano $n_n \sim p(n_n) = N(n_n | 0, \sigma_n^2)$ es equivalente en la media \tilde{m}_N a la solución de mínimos cuadrados regularizados.

la predictiva

Para un nuevo dato (x_*, t_*) , la distribución predictiva referente a la salida t_* se puede calcular como:

$$p(t_* | x_*, t, w) = \int p(t_* | x_*, w) p(w | t) dw$$

$$p(t_* | t) = \int N(t_* | \phi(x_*)^\top \tilde{m}_N, \tilde{G}_N) N(w | \tilde{m}_N, \tilde{G}_N) dw$$

$$p(t_* | x_*, t, w) = N(t_* | \phi(x_*)^\top \tilde{m}_N^\top, \tilde{G}_N^\top + \phi(x_*)^\top \tilde{G}_N \phi(x_*)^\top)$$